Journal of Bioinformatics and Computational Biology
Vol. 15, No. 1 (2017) 1750001 (16 pages)

DOI: 10.1142/S0219720017500019

New multilocus linkage disequilibrium measure for tag SNP selection

Bo Liao*, Xiangjun Wang, Wen Zhu, Xiong Li, Lijun Cai and Haowen Chen

College of Information Science and Engineering, Hunan University, Changsha, Hunan 410082 China
**dragonbw@163.com*

Received 9 February 2015
Revised 19 November 2015
Accepted 10 January 2017
Published 22 February 2017

Numerous approaches have been proposed for selecting an optimal tag single-nucleotide polymorphism (SNP) set. Most of these approaches are based on linkage disequilibrium (LD). Classical LD measures, such as D′ and r^2, are frequently used to quantify the relationship between two marker (pairwise) linkage disequilibria. Despite of their successful use in many applications, these measures cannot be used to measure the LD between multiple-marker. These LD measures need information about the frequencies of alleles collected from haplotype dataset. In this study, a cluster algorithm is proposed to cluster SNPs according to multilocus LD measure which is based on information theory. After that, tag SNPs are selected in each cluster optimized by the number of tag SNPs, prediction accuracy and so on. The experimental results show that this new LD measure can be directly applied to genotype dataset collected from the HapMap project, so that it saves the cost of haplotyping. More importantly, the proposed method significantly improves the efficiency and prediction accuracy of tag SNP selection.

Keywords: Tag SNP; linkage disequilibrium (LD); clustering algorithms; entropy.

1. Introduction

Single-nucleotide polymorphisms (SNPs) are important for genome-wide associations. In particular, recent developments in next-generation sequencing technology raised post-genomic studies to another level by considering the consideration of the contribution of rare SNP variants.[1,2] The challenge is no longer related to the generation of more data but rather focuses on the way how these genetic data can be efficiently analyzed to obtain sufficient power to detect and explain association signals. Millions of SNPs are present in the human genome. The number of SNPs is a challenge in studying complex diseases. Genotyping all SNP markers in the involved genomes are preferred, but this process is expensive and unnecessary. Given the existence of thousands of human SNPs with linkage disequilibrium (LD), researchers can select a small number of characteristic (tag) SNPs to represent the remaining

SNPs. The existence of correlation among SNPs may make a small fraction of the tag SNPs to be sufficiently useful. This small fraction of SNPs enables the inference of all other SNPs. The problem with tag SNP selection is the need to identify the smallest possible set of tag SNPs that would enable the precise inference of all other SNPs. A small tag SNP set obviously entails low genotyping cost.[3] Driven by such a significant potential benefit, various algorithms have been developed to select tag SNPs effectively. Such selection is essentially a feature selection problem from a machine-learning standpoint.

The available methods for tag SNP selection can be classified into two categories: haplotype-block-based methods and genome-wide methods.[4] The haplotype-block-based methods focus on the haplotype patterns in a population. This method assumes that the whole chromosome can be divided into blocks, which are separated by recombination hotspots, such that each block comprises a few recombination hotspots. Thus, tag SNP selection aims to identify the smallest possible tag SNPs for each block. In this way, all possible haplotype patterns in the block can be fully represented by the haplotype formed by the selected tag SNPs. However, no general solution exists for dividing the chromosome into blocks. Moreover, the lack of inter-block association degrades selection accuracy. By contrast, genome-wide methods do not divide a chromosome into blocks. These methods instead consider the correlation among the SNP markers across the entire genome to represent genome-wide associations (measured by pairwise LD). However, these genome-wide methods may cause the loss of some important information contained in the remaining SNPs and may thus fail to distinguish all the haplotypes in a LD cluster.

Numerous studies have been conducted to characterize LD patterns and to apply these patterns to tag SNP selection.[5–8] The most commonly used methods are based on pairwise LD measures, such as D′ and r^2. Despite their popularity, these measures assume that the individual haplotype phase has been previously resolved.[9–13] Given these technological limitations, most sequencing techniques provide genotype rather than haplotype information.[14,15] To satisfy the requirement of these LD measures for tag SNP selection, the haplotype phase should be estimated to computed r^2 for each pair of SNPs in the region, so that the haplotype information can be inferred from the genotype data. Although many algorithms can be used for haplotype inference[16–18] such as HAPLOTYPER,[19,20] PHASE[21–23] and so on, computational cost is extremely high, and most of these algorithms use statistical approaches [e.g. expectation–maximization (EM) algorithm] or machine-learning methods.[24–28] Although effective, many of these methods cause error rate defined in Refs. 19 and 21 of haplotype phase inference.[24] The accuracy of haplotype inference depends on several factors, including sample size, allele frequency, number of SNPs, missing data, and linkage between these SNPs. Thus, in order to avoid haplotype inference error and reduce computational cost, we propose a novel LD measure that can directly quantify the LD relationship between two SNP sets based on genotype data. Three scenarios are used to consider the LD among SNP markers. (1) One-to-one: considering two markers A and B, we can calculate the common pairwise LD measure r^2 between A and B to evaluate the linkage degree. When the value of r^2 exceeds a

given threshold (generally 0.8), the linkage between A and B is considered as a strong LD. In order to save on genotyping cost, we can select either A or B for further study. The selected marker is then considered as the tag SNP, whereas the other marker is the tagged SNP. (2) Many-to-one: three markers A, B, and C, are considered. If A and B are known to have a strong LD, then they are viewed as a whole. Thus, we introduce a new marker M that combines A and B. Furthermore, we determine the LD strength between C and M. r^2 is thus unsuitable for this case. Hao *et al.*[24] extended r^2 statistics to describe the statistical correlation between a group of markers (e.g. two or three) and another marker, which may solve this problem. However, this method consumes a great deal of time and memory. (3) Many-to-many: consider two SNP sets $S_1(A_1A_2, \ldots, A_n)$ and $S_2(B_1B_2, \ldots, B_m)$, where $n \geq 2, m \geq 2$. We determine whether the S_1 SNPs have strong LDs relative to those in S_2. We also determine methods by which to measure the LD between the S_1 and S_2 directly.

2. Materials and Methods

In this section, we first propose a new LD measure called the average information gain ratio (AIGR) on the basis of information theory. AIGR can measure the LD relationship between two SNP sets. This measure is a multilocus LD measure, which can also be directly applied to genotype data for tag SNP selection. Thus, we propose a novel SNP selection method called AIGR-Tagger which is used for tag SNP selection. The main ideas of the proposed approach are as follows: (1) SNP clustering: the SNPs are divided into different clusters, such that SNPs within the same cluster have strong LD (according to a given threshold). The clustering similarity measure is taken as the new LD measure. (2) Tag SNP selection: after SNP clustering, we select a tag SNP to represent each cluster. (3) Tag SNP evaluation: numerous approaches are used for tag SNP selection. However, most of these methods only focus on the number, rather than the quality of tag SNPs. The purpose of tag SNP selection is to choose practically useful SNPs that can best retain the allele information of all the SNPs in the candidate region. Thus, the tag SNP selection method is expected to select a small number of SNPs with minimal information loss. The selected tag SNPs are used to predict the unselected SNPs to develop efficient algorithms for selecting a proper set of SNPs with high prediction accuracy. We also introduce how the new LD measure value can be calculated between SNPs for both haplotype and genotype data.

2.1. *Information theory*

The following considerations will be based on Shannon entropy, which is denoted as $H(X)$, as a measure of genetic diversity and association. Considering a locus X with k alleles of frequency $p(x_i)(i = 1 \cdot k)$, the uncertainty of the random variable X is measured by

$$H(X) = -\sum_{i}^{k} p(x_i)\log_2 p(x_i). \tag{1}$$

The maximum $H(X)$ is achieved if all states are equally probable. In our association study, the joint entropy of loci X and Y can be defined as the entropy of the corresponding haplotype or genotype frequencies.

$$H(X,Y) = -\sum_i \sum_j p(x_i, y_j) \log_2 p(x_i, y_j). \tag{2}$$

The property of $H(X,Y)$ is that $H(X,Y) \leq H(X) + H(Y)$. Given this property, we determine how much information we can obtain from X if such loci is determined before Y. The conditional entropy $H(Y|X)$ of Y, given X, is equal to the proportion of the entropy of Y that remains after determining X, that is, $H(Y|X) = H(X,Y) - H(X)$. Therefore, the mutual information $I(X;Y)$ is the proportion of the entropy of Y that is removed after determining X.

$$I(X;Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X, Y) \tag{3}$$

Figure 1 clearly illustrates the mutual information $I(X;Y) = I(Y;X)$. We therefore unify the mutual information by using $I(X;Y)$ to represent such data. The correlation between two loci can be properly measured on the basis of their mutual information $I(X;Y)$. A large value of $I(X;Y)$ indicates high correlation. Notably, $0 \leq I(X;Y) \leq \infty$. However, we use $I(S_1; \mathrm{SNP}_x)$ and $I(S_2; \mathrm{SNP}_x)$ to represent the respective LD value of SNP sets S_1 and S_2 (S_1 and S_2 for two SNP sets exhibit high correlation among SNPs within each set) with another SNP_x. If the numbers of SNPs in S_1 and S_2 are unequal, $I(S_1; \mathrm{SNP}_x)$ and $I(S_2; \mathrm{SNP}_x)$ cannot be directly compared because a larger number set will result in greater mutual information with other SNPs. The key issue for this problem is the set scale because the mutual information cannot be compared if the set scale varies.

$$H(S_1, S_2) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(s_{1_i}, s_{2_j}) \log_2 p(s_{1_i}, s_{2_j}). \tag{4}$$

In Eq. (4), S_1 and S_2 contain one or more SNPs with n and m types. For example, S_1 contains $\{A, B\}$ with four types $\{AB, Ab, aB, ab\}$, with $n = 4$. S_2 contains $\{C, D\}$ with four types $\{CD, Cd, cD, cd\}$, with $m = 4$. In particular, S_1 or S_2 only have one

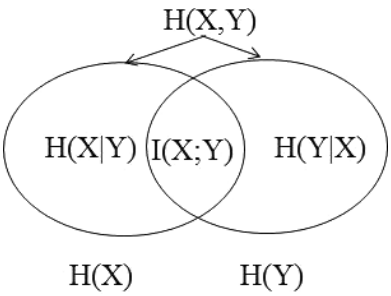

Fig. 1. Relationship between entropy and mutual information of X and Y.

SNP, with n or m equal to 2. $p(s_{1_i}s_{2_j})$ is the frequency of S_1S_2, like $ABCD$. Thus, $H(S_1, \text{SNP}_x)$ is a special form of $H(S_1, S_2)$ when S_2 contains only one SNP.

To make the scenario clearer, let us consider the following simple example in Table 1 which shows a dataset consisting nine SNPs. As follows, we provide a numerical concrete example for computing these three quantities, i.e. $H(X)$, $H(X, Y)$ and $I(X; Y)$.

$$H(\text{SNP1}) = -(0.7 * \log_2 0.7 + 0.3 * \log_2 0.3) = 0.8813,$$

$$H(\text{SNP2}) = -(0.6 * \log_2 0.6 + 0.4 * \log_2 0.4) = 0.97095,$$

$$H(\text{SNP1}, \text{SNP2}) = -(0.6 * \log_2 0.6 + 0.3 * \log_2 0.3 + 0.1 \log_2 0.1) = 1.2955,$$

$$I(\text{SNP1}; \text{SNP2}) = H(\text{SNP1}) + H(\text{SNP2}) - H(\text{SNP1}, \text{SNP2}) = 0.55675.$$

2.2. *New LD measure AIGR*

To address this problem, a new multilocus LD measure on the basis of AIGR is proposed. This measure is applied to calculate the LD between SNPs as shown in Fig. 2.

$$\text{AIGR}(S_1, S_2) = \frac{1}{2}\left(\frac{I(S_1; S_2)}{H(S_1)} + \frac{I(S_1; S_2)}{H(S_2)}\right). \tag{5}$$

In Eq. (5), $I(S_1; S_2)$ denotes the mutual information between S_1 and S_2, whereas $H(S_1)$ and $H(S_2)$ represent the entropy of S_1 and S_2, respectively. The property of Eq. (5), as discussed earlier, is summarized as follows: $0 \leq$ AIGR ≤ 1, with the lower and upper bounds attained when the SNPs are in complete linkage equilibrium (LE) and complete LD, respectively. This property is in good agreement with the mutual information and AIGR is a gain-ratio that can compare different SNP set scales. Notably, the LD of pairwise SNPs can also apply to this measure, which is a special case of AIGR. A large AIGR value indicates a strong correlation among SNP sets. The proposed measure overcomes the drawbacks of r^2 statistics, which can only be applied to haplotype data. Moreover, mutual information can only be used to measure the LD of pairwise SNP sets

Table 1. A simulated dataset.

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
$h1$	0	0	0	0	0	1	1	0	0
$h2$	1	1	0	1	1	0	0	0	0
$h3$	0	0	1	0	0	0	0	0	1
$h4$	0	0	0	0	0	0	0	1	0
$h5$	1	1	1	1	1	1	1	0	0
$h6$	0	0	0	0	0	0	0	1	0
$h7$	1	1	0	1	1	0	0	0	1
$h8$	0	0	0	0	0	1	1	1	1
$h9$	0	1	1	0	1	1	1	0	0
$h10$	0	0	1	0	0	0	0	1	0

with the same scale. In particular, the measure is computationally efficient and can thus handle any number of SNPs.

2.3. *Data coding*

In SNP selection, raw SNP data are usually converted to a matrix form. Assume that given n SNP sequences, each consisting of m SNPs. In this work, we are only interested in bi-allelic SNPs (i.e. SNPs taking only two different nucleotides among $\{a, g, c, t\}$ at the SNP position). Each haplotype can be represented by a binary string. The n sequences can form a matrix M of size $n * m$, where rows are sequences and columns are SNPs. Assume that no data are missing in the sequences. When these sequences are phased haplotypes, $M[i, j] \in \{0, 1\}$ represents the allele of the ith sequence at the jth SNP locus, where 0 and 1 pertain to the major and minor alleles, respectively. When these sequences are unphased genotypes, $M[i, j] \in \{0, 1, 2\}$, where 0 and 2 are homozygous types, which represent the major and minor alleles, respectively, and 1 indicates the heterozygous type.

2.4. *SNP clustering*

Hierarchical clustering algorithms are often used for their simplicity and efficiency. For tag SNP selection, the objects for clustering are SNPs. In this study, we apply our new LD measure called AIGR as an indicator of cluster similarity. In clustering, the two clusters with the largest AIGR that exceed the given threshold are merged into one cluster. Within the context of SNP clustering, a cluster should contain at least one SNP. At the beginning, for each SNP that belongs to an SNP cluster, a set S of SNPs, $\mathrm{S} = \{\mathrm{SNP}_1, \mathrm{SNP}_2, \ldots, \mathrm{SNP}_n\}$ and C of SNP clusters, $\mathrm{C}_i = \{\mathrm{SNP}_i\}$, $\mathrm{C} = \{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_m\}$. We then perform the calculation for every cluster with all other clusters $\mathrm{AIGR}(C_i C_j)$, if AIGR meets the following condition:

$$\max_{i,j}\{\mathrm{AIGR}(C_i, C_j)\} > \theta, \tag{6}$$

where θ represents a threshold value of AIGR. Moreover, when AIGR exceeds the given threshold θ (generally $\theta \in \{0.5, 1\}$; here, we set $\theta = 0.7$), then C_i and C_j will

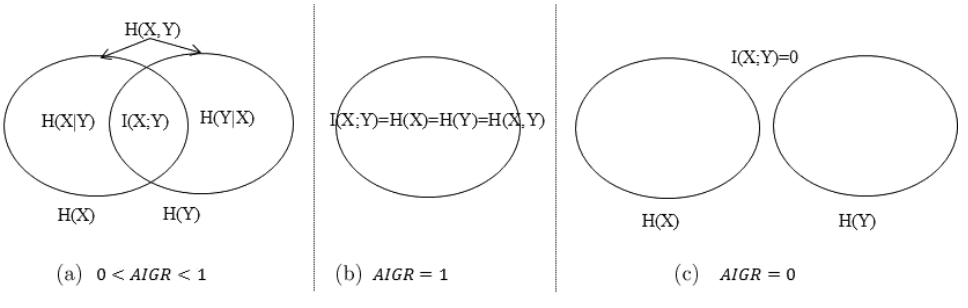

Fig. 2. AIGR relationship of X and Y. (a) General LD relationship of X and Y. (b) Complete LD of X and Y. (c) Complete LE of X and Y.

merge into $C'_i = C_i \cup C_j$. The clustering algorithm stops once the maximum AIGR is less than the given threshold θ or when the clusters merge into a single cluster. After clustering, we obtain k SNP clusters $C' = \{C'_1, C'_2, \ldots, C'_k\}$, $1 \leq k \leq m$, where k is the number of SNP clusters. Thus, within every SNP cluster, all SNPs have a strong LD with one another, which enables the selection of one SNP that can best represent the other SNPs as the tag SNP of the cluster.

2.5. *Tag SNP selection*

After SNP clustering, we identify k SNP clusters from which we select tag SNPs. The main idea is that when an SNP cluster has r SNPs, the tag SNP should best represent the SNP cluster, such that select the SNP that meets the following condition:

$$\max_i \frac{1}{r} \sum_{j=1, j\neq i}^{r} \mathrm{AIGR}(\mathrm{SNP}_i, \mathrm{SNP}_j). \tag{7}$$

In particular, if a cluster only has one SNP, we directly select this SNP as the tag SNP, because no other SNP can strongly correlate to the selected tag SNP. Every tag SNP can best represent the corresponding SNP cluster when this procedure is employed.

2.6. *SNP prediction*

After selecting the tag SNPs, the tag SNP sets undergo evaluation. Several methods can be used to assess a tag SNP selection method. Stram *et al.*[25] proposed a quality measure r^2, which is based on a subset of tag SNPs to predict the unselected SNP. This measure requires diploid data to infer haplotype from genotype data, and is thus inappropriate for our purpose. Carlson *et al.*[26] proposed a measure based on the accuracy of haplotype diversity, which is defined as the total number of bases among different haplotypes on the respective positions. The difference between couples of haplotypes pertains to the total number of differences among all SNPs. This measure, which was proposed by Carlson *et al.*,[26] can efficiently define the capability of tag SNPs to identify different haplotypes. However, this measure can only be used for haplotype blocks with limited haplotype diversity and is unsuitable for large datasets with multiple haplotype blocks and genotype data. Evaluating the capability of the tag SNP selection algorithm requires the prediction accuracy of tag SNPs for unselected SNPs. Cross-validation, such as leave-one-out cross-validation (LOOCV), is typically used for prediction. We predict each unselected SNP using LOOCV separately, to determine the average prediction accuracy. In addition, support vector machine (SVM) is used as the model.[27] Due to the good generalization ability and fast convergence rate, we apply radial basis function as kernel function and parameter c is set 100, gamma is set 0.1. Two strategies are used to access the prediction accuracy: (1) all tag SNPs collaboratively predict every unselected SNP, and (2) each tag SNP predicts the other SNPs within the corresponding cluster, and then average

all of the obtained SNPs. The second strategy is evidently better than the first for our method. Given n samples with m SNPs, we obtain k SNP clusters after SNP clustering. For every cluster, we select one SNP to represent all other SNPs within this cluster. To evaluate methods for tag SNP selection, these tag SNPs are used to predict every untagged SNP in the cluster, separately. LOOCV is applied to the samples. We use one sample as the test set and the remaining samples as training set. During training process, every SNP will be predicted by tag SNPs and then compared to their original value.

2.7. *Compression ratio*

To evaluate various methods comprehensively, we should assess the proportion of tag SNPs in the total SNP set. All tag SNP selection methods aim to select an optimal tag SNP set, which contains the smallest possible number of SNPs but with the least possible information loss. In this study, compression ratio is used to measure this factor. Compression ratio is defined as:

$$\text{compression ratio} = \frac{N_{\text{tag}}}{N_{\text{total}}}. \tag{8}$$

In Eq. (8), N_{tag} represents the number of selected tag SNPs, and N_{total} represents the number of total SNPs. A smaller the compression ratio indicates a better tag SNP selection method.

2.8. *Results*

In this section, we implemented the AIGR-Tagger in C++ and conducted experiments on a Pentium 4 processor with a 2-GB RAM running on Windows 7. To evaluate the properties and performances of AIGR-Tagger, our tag SNP selection method is tested on both haplotype and genotype data. All datasets were obtained from HapMap.[28] The detailed information of 10 ENCODE region datasets used in our study is shown in Table 2. The number of individuals with genotype (haplotype)

Table 2. Datasets of ENCODE region SNPs.

Region name	Chromosome band	Genomic interval	Genotype SNP numbers (MAF > 1%)	Haplotype SNP numbers (MAF > 1%)	Genotyping group
ENm010	7p15.2	Chr7:26924045..27424045	756	322	UCSF-WU, Perlegen
ENm013	7q21.13	Chr7:89621624..90121624	1053	483	Broad, Perlegen
ENm014	7q31.33	Chr7:126368183..126865324	1135	611	Broad, Perlegen
ENr112	2p16.3	Chr2:51512208..52012208	1273	751	McGill-GQIC, Perlegen
ENr113	4q26	Chr4:118466103..118966103	1401	772	Broad, Perlegen
ENr123	12q12	Chr12:38626477..39126476	1312	772	BCM, Perlegen
ENr131	2q37.1	Chr2:234156563..234656627	1335	835	McGill-GQIC, Perlegen
ENr213	18q12.1	Chr18:23719231..24219231	882	529	Illumina, Perlegen
ENr232	9q34.11	Chr9:130725122..131225122	742	390	Illumina, Perlegen
ENr321	8q24.11	Chr8:118882220..119382220	877	484	Illumina, Perlegen

data for each region is 90(120). The genotype SNP numbers is unequal to the haplotype SNP number because some information is lost when the haplotype phase is inferred from genotype data.

For comparison, two other existing methods in the literature are used to analyze the same haplotype SNP datasets. These methods are: (1) Fast Tagger, which is based on multi-marker LD.[13] and (2) Feature Selection and Feature Similarity (FSFS).[29] However, the two methods only selected the SNPs but did not evaluate the effectiveness of these selected tag SNPs. Thus, LOOCV is used to evaluate the performance of the three methods. All methods aim to select a set of highly predictive tag SNPs for unselected SNPs. Therefore, the selected number of tag SNPs, prediction accuracy and compression ratio are used as performance criteria to evaluate all three methods. The performances of these three methods on each ENCODE region are summarized in Table 3 and Fig. 3.

We also compute the overlapping ratio of tag SNPs selected by AIGR-Tagger and Fast Tagger. In this study, the overlapping ratio of the two tag SNP selection methods can be computed using Eq. (9).

$$\text{overlapping ratio} = \frac{2 * (N_{\text{tag1}} \cap N_{\text{tag2}})}{N_{\text{tag1}} + N_{\text{tag2}}}, \tag{9}$$

where N_{tag1} and N_{tag2} denote the number of selected tag SNPs using methods one and two, respectively. Since can only get the number of selected tag SNPs by FSFS, we cannot compute for the overlapping ratio of tag SNPs selected by FSFS and the other two methods. The result of the overlapping ratio of tag SNPs selected by AIGR-Tagger and Fast Tagger is shown in Table 4.

For comparison, the same 10 ENCODE regions on genotype datasets are used to analyze the proposed method, AIGR-Tagger. The performance of AIGR-Tagger on genotype datasets is shown in Table 5.

Table 3. Summary of the number of selected tag SNPs, prediction accuracy and compression ratio on the haplotype datasets of AIGR-Tagger, Fast Tagger and FSFS.

		Number of tag SNPs			Prediction accuracy			Compression ratio		
Region	SNP numbers (MAF > 1%)	AIGR-Tagger	Fast Tagger	FSFS	AIGR-Tagger	Fast Tagger	FSFS	AIGR-Tagger	Fast Tagger	FSFS
ENm010	322	**79**	95	110	**0.9902**	0.9574	0.9489	**0.2453**	0.2950	0.3416
ENm013	483	**37**	57	92	0.9689	**0.9900**	0.9887	**0.0766**	0.1180	0.1905
ENm014	611	**72**	101	172	**0.9846**	0.9755	0.9408	**0.1178**	0.1653	0.2815
ENr112	751	**83**	114	230	**0.9820**	0.9581	0.9429	**0.1105**	0.1518	0.3063
ENr113	772	**70**	96	175	**0.9710**	0.9698	0.9592	**0.0907**	0.1244	0.2267
ENr123	772	**88**	120	223	**0.9784**	0.9720	0.9599	**0.1140**	0.1554	0.2889
ENr131	835	**124**	153	310	**0.9829**	0.8983	0.8567	**0.1485**	0.1832	0.3713
ENr213	529	**72**	86	185	**0.9915**	0.9726	0.9607	**0.1361**	0.1626	0.3497
ENr232	390	**69**	86	133	**0.9833**	0.9491	0.9230	**0.1769**	0.2205	0.3410
ENr321	484	**72**	88	128	**0.9806**	0.9690	0.9591	**0.1488**	0.1818	0.2645
Overall	5949	**766**	996	1758	**0.9813**	0.9612	0.9440	**0.1288**	0.1674	0.2955

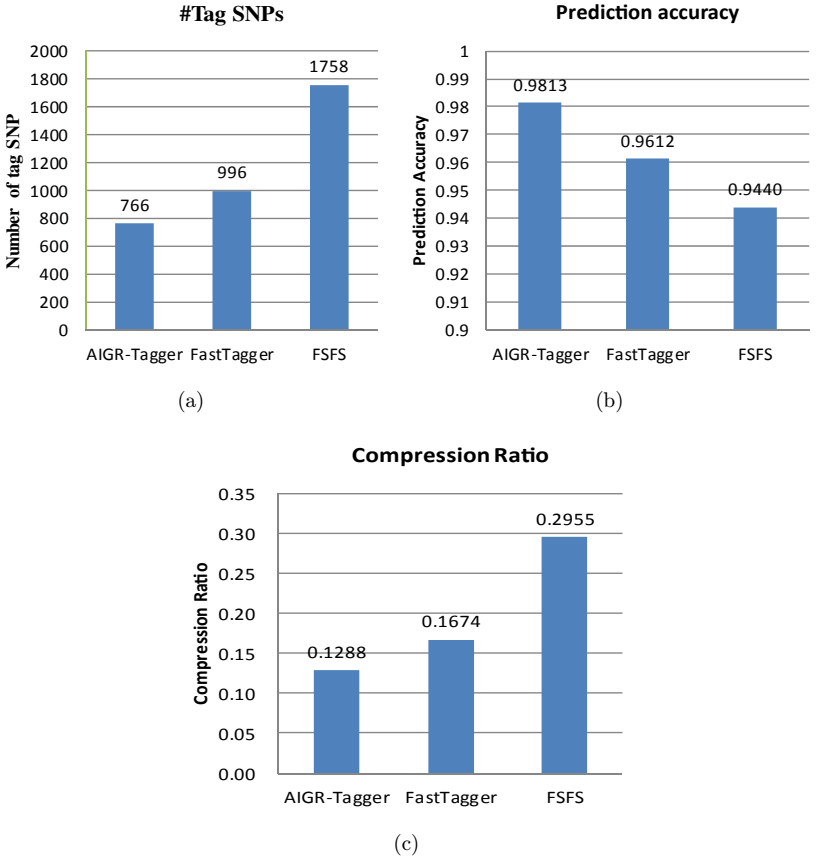

Fig. 3. Comparison of selected tag SNPs numbers, prediction accuracy and compression ratio of the different methods on the haplotype datasets. (a) Number of selected Tag SNPs. (b) Prediction accuracy. (c) Compression ratio.

Table 4. Summary of the overlapping ratio of tag SNPs selected by AIGR-Tagger and Fast Tagger.

		Number of tag SNPs		
Region	SNP number (MAF > 1%)	AIGR-Tagger	Fast Tagger	Overlapping ratio
ENm010	322	79	95	0.7356
ENm013	483	37	57	0.5106
ENm014	611	72	101	0.7399
ENr112	751	83	114	0.5990
ENr113	772	70	96	0.6145
ENr123	772	88	120	0.6154
ENr131	835	124	153	0.7292
ENr213	529	72	86	0.7595
ENr232	390	69	86	0.7226
ENr321	484	72	88	0.7250
Overall	5949	766	996	0.6751

Table 5. Summary of the performance of AIGR-Tagger on the genotype datasets.

Region	SNP numbers (MAF > 1%)	Number of tag SNPs	Prediction accuracy	Compression ratio
ENm010	756	246	0.9829	0.3254
ENm013	1053	151	0.9682	0.1434
ENm014	1135	238	0.9891	0.2097
ENr112	1273	280	0.9781	0.2200
ENr113	1401	254	0.9824	0.1813
ENr123	1312	274	0.9886	0.2088
ENr131	1335	334	0.9852	0.2502
ENr213	882	218	0.9881	0.2472
ENr232	742	220	0.9847	0.2965
ENr321	877	219	0.9749	0.2497
Overall	10766	2434	0.9822	0.2332

We also compared the properties and performances of AIGR-Tagger and the commonly used software Haploview.[30] Haploview is the most commonly used software for SNP analyses. Haploview is designed to simplify and expedite the process of haplotype analysis by providing a common interface for several tasks relating to such analyses. This software can also be used to select tag SNPs based on haplotype data. The performance of Haploview selecting tag SNPs on the haplotype datasets is summarized in Table 6. Notably, the threshold of r^2 is set as 0.8. Table 6 shows that the SNP Numbers of the ENCODE Region are unequal to the corresponding Haplotype SNP Numbers in Table 2. Given that Haploview has a specific dataset download module to obtain SNP datasets; it excludes individuals in the process of online download. Thus, Haploview cannot determine the prediction accuracy on the basis of the selected tag SNPs. For comparison, we focus the compression ratio of Haploview and of the other methods.

In addition to these results, we also compare our method with another entropy-based method called ER and these results are listed in Table 7.

Table 6. Summary of the performance of Haploview on the haplotype datasets.

Region	SNP numbers	Number of tag SNPs	Compression ratio
ENm010	231	108	0.4675
ENm013	257	84	0.3268
ENm014	298	114	0.3826
ENr112	278	122	0.4388
ENr113	287	120	0.4181
ENr123	307	117	0.3811
ENr131	547	215	0.3931
ENr213	221	100	0.4525
ENr232	230	105	0.4565
ENr321	265	112	0.4226
Overall	2921	1197	0.4098

Table 7. The comparison between two entropy-based methods.

Region	Number of tag SNPs by AIGR-Tagger	Number of tag SNPs by ER
ENm010	246	287
ENm013	151	159
ENm014	238	238
ENr112	280	299
ENr113	254	231
ENr123	274	302
ENr131	334	354
ENr213	218	202
ENr232	220	228
ENr321	219	220

Table 7 shows that our method selects smaller subset of tag SNPs than ER. It indicates that our method can select more representative tag SNPs, so that the cost of genotyping is significantly saved.

3. Discussion and Conclusion

Classical LD measures, such as D' and r^2, are often used to measure the degree of LD between two loci. However, these measures fail to provide a direct measure of joint LD among multiple loci. Meanwhile, most of these measures require haplotype data. Owing to technological limitations, most sequencing techniques provide genotype, rather than haplotype data. Notably, our "real" haplotype distributions are estimated from the genotype data. Thus, the haplotype phase inferred from genotype data is only an estimate, the accuracy of which is dependent on the amount of available data. In this study, we propose an effect multilocus LD measure on the basis of AIGR to overcome these problems. Our LD measure can quantify the extent of LD between two SNP sets and can also be directly applied to genotype data for tag SNP selection. This measure considers the interactions among SNPs and may be beneficial for follow-up studies, such as epistasis analysis.

On the basis of AIGR, a tag SNP selection algorithm called AIGR-Tagger is proposed for both haplotype and genotype data. We compared the properties and performances of our proposed method with other two state-of-the-art tag SNP selection methods, namely, Fast Tagger and FSFS, on haplotype datasets. In Table 2, AIGR-Tagger selected less tag SNPs but achieved higher prediction accuracy than the other two methods. Even with the same total number of SNPs used, fewer tag SNPs the result in a smaller compression ratio. This result shows that AIGR-Tagger can select less SNPs to represent the other unselected SNPs better. On ENm013, AIGR-Tagger selected 37 tag SNPs to achieve 96.89% prediction accuracy, this prediction accuracy, which is slightly lower than that of Fast Tagger. Thus, we believe that AIGR-Tagger outperforms Fast Tagger on ENm013 because the former lost only little minimal information while achieving a high degree data of compression. Table 3 shows that the tag SNP overlapping ratio of AIGR-Tagger and Fast

Tagger is nearly 70%. This result indicates that the selected tag SNPs between these two methods have a high degree overlapping. Given the complete LD among SNP loci, we believe that the actual overlapping ratio is higher than the value in Table 3. Obviously, our method is a block-free tag SNP selection method, which can address two issues of block-based: the inconsistency of definition of the block and computational cost of block partition.

The running time is a measure to evaluate the efficiency of an algorithm. Usually, the running time of an algorithm is stated as a function relating the input length to the number of steps (time complexity) or storage locations (space complexity). However, the running time measure is too specific to compare different algorithms fairly for the following reasons: (1) the running time relies on the specific machine. (2) as the size of datasets changes, the running time may change dramatically. A simple running time lasting only seconds is difficult to use for describing the efficiency under different situations. (3) given that algorithms are platform-independent (i.e. a given algorithm can be implemented in an arbitrary programming language on an arbitrary computer running an arbitrary operating system), significant drawbacks exist relative to using an empirical approach to gauge the comparative performance of a given set of algorithms. Consequently, we use big O notation to denote the time complexity of an algorithm. Suppose we have n samples with m SNPs. The whole computational complexity of Fast Tagger is $O(n * m^k)$. In this work, k represents the number of SNPs in a tagging rule in Fast Tagger. k is set 2 or 3 because calculating multi-marker r^2 statistics is more expensive than computing for pairwise r^2. The overall computational complexity of FSFS is $O(k * m * n^2)$, in which k is a parameter of KNN in FSFS. The computational complexity of AIGR-Tagger is $O(n * m^2)$. AIGR-Tagger is evidently faster than the other two methods.

As shown in Table 4, AIGR-Tagger can select tag SNPs on genotype data. For comparison, an experiment is conducted on the same ENCODE region with haplotype datasets. Given the information losses that occur when the haplotype phase is inferred from genotype data, the numbers of SNPs are unequal. Table 4 shows that AIGR-Tagger can effectively select tag SNPs on genotype data. This property is very useful. If only genotype data are acquired, tag SNPs can be selected without haplotype phasing from the genotype data. Thus, AIGR-Tagger can reduce the cost of haplotype phasing and avoid the information losing. At the same time, AIGR-Tagger is compared with software Haploview based on haplotype data. Through the comparison shown in Tables 2 and 5, we find that AIGR-Tagger selected only 12.88% tag SNPs from the total SNPs with 98.13% prediction accuracy. This comparison results shows that the data compression of AIGR-Tagger is better than that of Haploview with minimal information loss.

Consequently, the performance of current tag SNP selection methods is limited by certain restrictions such as the small bounded location or the fixed number of predictive tag SNPs. Moreover, most methods can only be applied to two markers (pairwise) LD or require an additional haplotype inference as pre-processing. Our goal is to address these limitations and improve the performance of currently

available tag SNP selection methods. That is, our method is neither limited to bi-allelic SNPs nor having an additional haplotype inference-procedure. Moreover, our method is based on information theory. AIGR is a multilocus LD measure which considers the LD between multiple loci that can further capture LD than bi-allelic measure, like r^2. This feature of AIGR-Tagger facilitates follow-up studies and improves the confidence of medical or biological researchers in bioinformatics.

Acknowledgments

This work is supported by the Program for New Century Excellent Talents in University (Grant Np. NCET-10-0365), National Nature Science Foundation of China (Grant Nos. 11171369, 61272395, 61370171, 61300128 and 61572178), the National Nature Science Foundation of Hunan Province (Grant No. 12JJ2041), the Planned Science and Technology Project of Hunan Province (Grant No. 2012FJ2012). And this study is also supported by Hunan Provincial Innovation Foundation for Postgraduate (CX2013A007), the Scholarship Award for Excellent Doctoral Student granted by Ministry of Education, China and supported by the Fundamental Research Funds for the Central Universities, Hunan University.

References

1. Wu C, Cui Y, Boosting signals in gene-based association studies via efficient SNP selection, *Brief Bioinform* 2013.
2. Shi Y *et al.*, Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome, *Nature Genet* **44**(9):1020–1025, 2012.
3. Johnson GC *et al.*, Haplotype tagging for the identification of common disease genes, *Nature Genet* **29**(2):233–237, 2001.
4. Wang W-B, Jiang T, A new model of multi-marker correlation for genome-wide tag SNP selection, *Proc Int Conf Genome Informatics*, World Scientific, 2008.
5. Patil N *et al.*, Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21, *Science* **294**(5547):1719–1723, 2001.
6. Zhang K *et al.*, A dynamic programming algorithm for haplotype block partitioning, *Proc Natl Acad Sci* **99**(11):7335–7339, 2002.
7. Nothnagel M *et al.*, Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks, *Human Hered* **54**(4):186–198, 2002.
8. Rinaldo A *et al.*, Characterization of multilocus linkage disequilibrium, *Genet Epidemiol* **28**(3):193–206, 2005.
9. Liu Z, Lin S, Multilocus LD measure and tagging SNP selection with generalized mutual information, *Genet Epidemiol* **29**(4):353–364, 2005.
10. Qin ZS, Gopalakrishnan S, Abecasis GR, An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria, *Bioinformatics* **22**(2):220–225, 2006.
11. Chen Y-H, Chen T, An Integer programming approach for the selection of tag SNPs using multi-allelic LD, *Commun Inf Syst* **9**(3):253–268, 2009.
12. Chuang L-Y, Hou Jr Y-J, Yang C-H, A Novel prediction method for tag SNP selection using genetic algorithm based on KNN, *Int J Chem Biomol Eng* **3**(1), 2010.

13. Liu G, Wang Y, Wong L, FastTagger: An efficient algorithm for genome-wide tag SNP selection using multi-marker linkage disequilibrium, *BMC Bioinform* **11**(1):66, 2010.
14. Chuang L-Y, Huang W-L, Yang C-H, An improved particle swarm optimization for tag single nucleotide polymorphism selection, *Proc Int Multi Conference of Engineers and Computer Scientists*, 2012.
15. Chen WP, Hung CL, Lin YL, Efficient haplotype block partitioning and tag SNP selection algorithms under various constraints, *Biomed Res Int* **2013**:984014, 2013.
16. Niu T, Algorithms for inferring haplotypes, *Genet Epidemiol* **27**(4):334–347, 2004.
17. Scheet P, Stephens M, A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase, *Am J Human Genet* **78**(4):629–644, 2006.
18. Browning SR, Missing data imputation and haplotype phase inference for genome-wide association studies, *Human Genet* **124**(5):439–450, 2008.
19. Niu T *et al.*, Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms, *Am J Human Genet* **70**(1):157–169, 2002.
20. Zhang Y, Niu T, Liu JS, A coalescence-guided hierarchical bayesian method for haplotype inference, *Am J Human Genet* **79**(2):313–322, 2006.
21. Stephens M, Smith NJ, Donnelly P, A new statistical method for haplotype reconstruction from population data, *Am J Human Genet* **68**(4):978–989, 2001.
22. Stephens M, Donnelly P, A comparison of bayesian methods for haplotype reconstruction from population genotype data, *Am J Human Genet* **73**(5):1162–1169, 2003.
23. Stephens M, Scheet P, Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation, *Am J Human Genet* **76**(3):449–462, 2005.
24. Hao K, Di X, Cawley S, LdCompare: Rapid computation of single-and multiple-marker r2 and genetic coverage, *Bioinformatics* **23**(2):252–254, 2007.
25. Stram DO *et al.*, Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study, *Human Hered* **55**(1):27–36, 2003.
26. Carlson CS *et al.*, Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans, *Nature Genet* **33**(4):518–521, 2003.
27. Chang C-C, Lin C-J, LIBSVM: A library for support vector machine, *ACM Trans Intell Syst Technol (TIST)* **2**(3):27, 2011.
28. Gibbs RA *et al.*, The international HapMap project, *Nature* **426**(6968):789–796, 2003.
29. Phuong TM, Lin Z, Altman RB, Choosing SNPs using feature selection, *Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE*, IEEE, 2005.
30. Barrett JC *et al.*, Haploview: Analysis and visualization of LD and haplotype maps, *Bioinformatics* **21**(2):263–265, 2005.

Bo Liao received the PhD degree in computational mathematics from the Dalian University of Technology, China, in 2004. He is currently a professor at Hunan University. He was at the Graduate University of Chinese Academy of Sciences as a postdoctorate from 2004 to 2006. His current research interest includes bioinformatics, data mining, and machine learning.

Xiangjun Wang received the MSc degree from the College of Information Science and Engineering, Hunan University, China. His research interests include data mining and tag SNPs selection.

Wen Zhu received the MSc degree in computer science and technology from Hunan University, China, in 2010, where she is currently a lecturer. Her current research interests include bioinformatics, data mining, and machine learning.

Xiong Li received the BSc degree from Xi'an Shiyou University, China, in 2009 and is currently working towards a PhD degree from the College of Information Science and Engineering, Hunan University, China. He is currently with the genome-wide association study. His research interests include data mining and genome-wide association studies.

Lijun Cai received the PhD degree in computer application technology. He is currently a professor of computer science and technology at Hunan University. His main research interests include cloud computing and bioinformatics.

Haowen Chen was born in Hunan, China. Currently, he is working in Hunan University. He received his MSc degree in computer science and technology from Hunan University, China. His current research interests include bioinformatics, data mining, and machine learning.